DATA MANAGEMENT AND SERVICES FOR GLOBAL CHANGE RESEARCH

Donald J. Collins, Susan Heinz

The DAAC Alliance

Abstract

The DAAC Alliance has undertaken an activity to describe best practices in data management. In the SEEDS era, as these principles are accepted, it is expected that they will be used to define data center requirements.

This paper will summarize the present findings of that effort, and will describe the DAAC Alliance's view of the fundamental concepts of data management and services to which successful data centers should subscribe.

This work is ongoing, and the input of the community is solicited to give this work the broadest applicability.

Introduction

The national data centers are the stewards of the global long-term climate record, whether held in active or long-term archives.

In the management of these archives, Federal policy mandates the full and open access to the full suite of quality data for global change research. To accommodate this policy, and to meet the needs of the global change research community, requires a continuing commitment to the establishment, maintenance, validation, description, accessibility, and distribution of high-quality, long-term data sets.

The Office of Science and Technology Policy, under the direction of D. Allan Bromley, issued a document in 1991 entitled "Data Management for Global Change Research Policy Statements" which included the following principles:

- Preservation of all data needed for long-term global change research is required. For each and every global change data parameter, there should be at least one explicitly designated archive. Procedures and criteria for setting priorities for data acquisition, retention, and purging should be developed by participating agencies, both nationally and internationally. A clearinghouse process should be established to prevent the purging and loss of important data sets.

- Data archives must include easily accessible information about the data holdings, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining the data.

- National and international standards should be used to the greatest extent possible for media and for processing and communication of global data sets.

- Data should be provided at the lowest possible cost to global change researchers in the interest of full and open access to data. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. Agencies should act to streamline administrative arrangements for exchanging data among researchers.

The NASA Earth Science Enterprise, in stating the Data Management Principles for the EOS Program, included the following data management principles:

- NASA is committed to the full and open sharing of Earth science data obtained from U.S. Government-funded and –owned systems with all users as soon as such data become available. All Earth System Enterprise missions, projects, grant proposals shall include data management plans to facilitate implementation of this principle.

- For data from Government-owned or –funded systems, NASA will enforce a principle of non-discriminatory access so that all users within the same data use category will be treated equally.

- All data required for long-term global change research shall be archived. Data archives shall include easily accessible information about the data holdings, including quality assessments, supporting relevant information, and guidance for locating and obtaining data.

    - All acquired EOS data will be processed at least to Level 1, and archived at Level 0 or at a higher level from which Level 0 may be recovered.

This paper is intended to address the present best practices for the stewardship of the global long-term climate record, and to recommend standards for data stewardship for the preservation of the nation's archive of these global data. It is anticipated that the application of these principles will lead to the cost-effective stewardship of Earth science data for use by future generations.

The principal challenges to the appropriate stewardship of the global long-term climate record are derived from the observation that the national holdings of global environmental data are increasing rapidly. This increase is illustrated in the following summary figure for the data holdings of NASA, NOAA and the USGS.

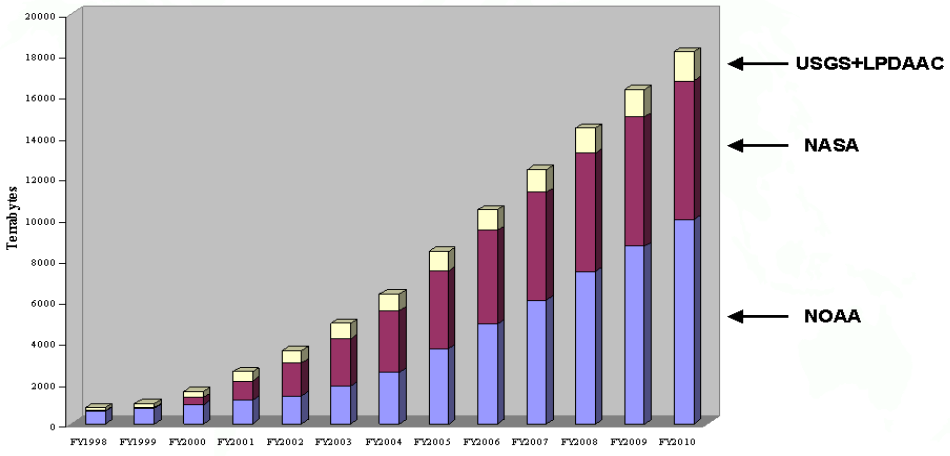

Figure 1: The projected increase in global climate data.

The increase in data volume anticipated from the NASA global change research program results from the significant number of NASA flight missions illustrated in the following figure.

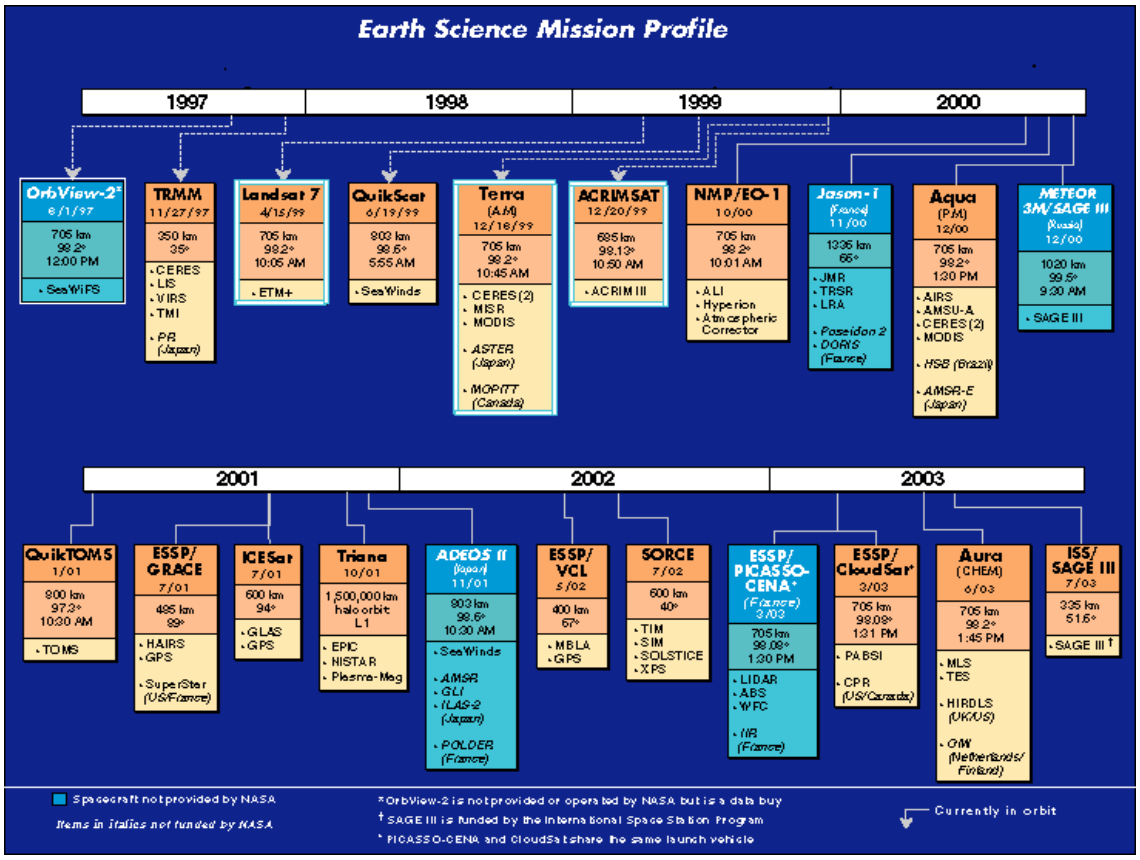

Figure 2: NASA Earth Science Missions

Because of the significant increase in data volumes, and in the increasing complexity of the data, the stewardship of these data is expected to become increasingly difficult, and the costs are expected to increase accordingly. The demand for services in data management will increase because of the growing national data holdings, and because of the demand to preserve longer temporal ranges of datasets and the need to compare data from multiple and cross-discipline datasets. These facts, coupled with the prospect that the budgets available for data stewardship are likely to remain at the present levels, leads to the conclusion of the necessity for increased efficiency in all aspects of data stewardship in the future. In particular, the role of technology in increasing efficiency, increasing the availability of services, and reducing cost is critical to the success of efforts to preserve the nation's archive of environmental data.

This paper will present a framework for the establishment of a national Earth science data management strategy that will be consistent with Federal Data Management Policy, and will reflect the policies of the agencies that are committed to the support of long-term global change research.

The goal of this work is to maximize the services available to the scientific community and to minimize the impact of budgetary considerations on the scientific community through improved and consistent data management principles and practices while protecting the data and information in the archives for access by future generations of Earth scientists, and by the general public.

Data Management Life Cycle

The DAAC Alliance has taken a 'full life cycle' approach to data stewardship. This approach includes the identification of the data to be archived, the planning for the data acquisition and archiving, and the data ingest into the data center. The full life cycle also includes data production and reprocessing, data archiving, data distribution, user services, and finally the disposition of data that is no longer of scientific interest. The DAAC Alliance recognizes that the stewardship of scientific data includes the availability of services for the user community. The level of these services will be addressed separately, but are clearly linked to the data management principles described in this work.

Identification of the data to be archived

The National Data Centers must strive to understand what data are required to form the global long-term climate record as a resource for future research on climate and global change.

The identification of these data must proactively involve the science community in all decisions about the selection and preservation of data and supporting information for future science use. The scientific community must also be involved in the definition of the life cycle of the data to be archived at the Data Centers. Thus, a central activity for the Data Centers is to identify the scientific communities to be served.

This requires that the National Data Centers understand both the needs of the National Earth science programs, and the needs of the broader user community. Understanding of the needs of the National Earth science programs is sought from two perspectives, the program scientists who define, fund and guide the programs, and the funded investigators whose work is a part of the programs. The individual investigators can validate the Data Center understanding of their needs for products and services, while the program scientists can provide a higher level view of the priorities and directions of the programs to guide data stewardship efforts.

To accomplish this, the National Data Centers must understand the data, products and services that the scientific users need most. This understanding includes an understanding of what makes a data set, product or service 'high quality' for users, and what forms or formats for data and products most effectively enable users' research and discovery. The Data Centers must also understand what data, products, and services are most useful for Earth science applications. These activities will become most effective to a wide range of users by seeking out and capitalizing on opportunities to support key scientific communities, as well as new users, and to collaborate on developing new applications.

The National Data Centers must preserve the data, and the knowledge about the data through the active use of the data by staff engaged in scientific research.

The National Data Centers must also understand the transient nature of the data that they manage. This includes an understanding of the versions of the algorithms used to produce the various levels of the data, and the maintenance of all versions of the data presently used in scientific research, and the maintenance of the algorithms, production software, and the documentation of those algorithms and software used to produce the various versions of the data.

The establishment of Data Center Science Advisory Groups will ensure that science priorities guide decisions as to which data is obtained and archived, and will help set priorities for data retention and advise on data disposition. These Science Advisory Groups will help the Data Centers to develop priorities based on science needs. To assist in this process, the Data Centers must provide for the periodic review of existing data holdings and prioritized recommendations for data to be added and for data identified as candidates for disposition.

Planning for data acquisition and archiving

The National Data Centers must provide cost-effective systems for managing the data.

Data Center Architecture and Infrastructure planning:

The National Data Centers must tailor the provision of data and information services to meet the specific needs of key Earth science disciplines, including multidisciplinary / interdisciplinary research and modeling. To address these issues adequately, the Data Centers must ensure that attention is given to cross-references of corroborative and ancillary data, which are increasingly important for interdisciplinary studies involving cross-sensor synergy, and that the full body of knowledge about the data and its generation is preserved. The scientific community must be involved in the key decisions that determine the services and performance to be provided by the Data Centers, including peer review of the Data Center performance.

The National Data Centers must implement an archive system to meet three major challenges; support of reprocessing of large, long-term data sets, migration to new media, and the distribution to users. The success of these activities depends upon the performance characteristics of the system, data organization within the archive, and the format of the data.

The Data Centers must define archive delivery performance standards to ensure that archive systems will support reprocessing of long-term data sets and media migration. These standards will include the organization of data within the

archives, and the configuration of the archive systems, to ensure effective support for data reprocessing and anticipated media migration.

The Data Centers must formulate and adopt standards and practices that promote superior data stewardship and facilitate the exchange of information between Data Centers to permit each Data Center to develop the best approach for its own archive systems. These standards include data access standards for metadata, read software, documentation, and the data format standards and the media formats to be supported for data distribution.

Data Management Plans:

The National Data Centers must actively assist NASA and other agency flight projects and field campaigns in planning their requirements for data products, data formats, read software and documentation. This will be accomplished through the participation in the generation of these plans, through the review of early versions of project documentation, and by providing useful tools to assist the project in complying with accepted standards. In the course of providing assistance, the Data Centers will secure the greatest possible degree of compliance with the accepted standards.

The National Data Centers must assist the flight projects in the development of the Project Data Management Plans to insure consistency in the standards of data stewardship between the Project and the Data Centers. These plans will provide a common understanding of the data to be managed, and permit the establishment of the most cost-effective data stewardship for the mission data.

Long-Term Archive planning:

The National Data Centers must provide for the permanent archiving of key data sets to support the global long-term climate record.

This activity includes data maintenance, and the maintenance of the data formats, documentation, read software, and the science data product generation software, to permit access to the data in the future.

Data ingest

The National Data Centers must provide secure systems for accessing the data from the flight projects and for the preservation of the data.

These systems will manage the flow of the data into the archive and provide data services that enhance the data as it is acquired by the Data Center.

Metadata:

These systems include the production of Metadata, including the generation of version control identification, and the versions of the production software required to recreate the data from the lower level data.

Data set completeness:

The Data Centers will insure the completeness of their data collections, including the archiving of the required ancillary data, project and data set documentation, and the science production software that will be needed by a scientist In the future to make full use of the data, without loss or degradation in quality. This includes a requirement for the preservation of the lowest levels of the data that are independently usable by the scientific community, and by the general public.

The data documentation must contain information needed to support the use of the data set in conjunction with other data sets.

Quality Assurance:

The National Data Centers must ensure that data ingested into the archives is of the highest quality. The Data Centers, through cooperation with the scientific community, will develop, advocate and implement standards and practices for data quality, and for quality monitoring, which assure the utility of the data within the scientific community. This activity embraces product generation and quality monitoring of ingest and distribution functions.

When data is provided by other than the flight projects, the data providers must be sufficiently engaged to ensure that science quality checked data is provided in an acceptable format with quality ancillary products and complete supporting documentation.

Data production and reprocessing

The National Data Centers must provide robust systems for processing the data from their archives.

To insure the scientific integrity of the data products, the National Data Centers must involve the scientific community in the definition and development of the algorithms used for processing science data products. These algorithms will be used to process the data from the lowest level satellite data to create the highest quality science data products.

The Data Centers will provide for the reprocessing of data to accommodate the latest scientific findings into the data products. The reprocessing of data

holdings must be accompanied by the documentation of the production characteristics, and the quality control of the production process.

Data Archiving

The National Data Centers must preserve the data, and preserve the knowledge about the data.

Archive management:

The National Data Centers must define and maintain science-approved standards and guidelines for data archiving.

These standards include data archive format standards that are consistent with community expectations and practices. These data formats must be both forward and backward compatible to insure access by the scientific community. The National Data Centers must optimize data access and convertibility to community standard distribution formats, including appropriate metadata standards.

The National Data Centers must provide for the maintenance of multiple versions of the data, and of the information about the data, as required by the scientific community. The decisions as to which data should be preserved must be made in coordination with the scientific community.

The National Data Centers must provide clear, concise descriptions of the data, its source and processing history. This documentation must include the scientific algorithms used to process the data and a description of the processing steps and ancillary data used in its production.

The National Data Centers must make available read software in programming languages that are in common use by the scientific community, and maintain these software packages as technology advances within that community.

The National Data Centers must provide for the regular, periodic backup of all data and information held in the archive to preserve the integrity of the data for future use.

The National Data Centers must ensure the migration of data to new media and new hardware and software technology to avoid having data trapped on unsupported or obsolete technology. The migration of data to new technologies will include the updating of metadata and data formats according to evolving community standards, and the automation of migration activities to reduce cost.

The migration of data to new technologies will be driven by media-refresh requirements or archive system replacements and upgrades. The Data Centers must plan for archive technology upgrade and insertion on a 3-5 year cycle, and must be continuously alert to the possibility of data sets suddenly becoming vulnerable. As the data is migrated, the Data Centers must ensure the quality and integrity of the data through active monitoring as appropriate for data protection. To enhance the utility of the data and information, the National Data Centers must avoid proprietary software environments that limit flexibility in system development and maintenance. The use of Open Source solutions may provide benefits of cost and flexibility.

Archive content:

Data
Metadata
Production S/W source code
S/W to read
Documentation on data, metadata and formats
Ancillary data
Cal/Val information and data
QA information

Deep Archive:

The Data Centers must insure the maintenance of an off-site backup copy of all data and information that is held in the archive.

Long-Term Archive:

The Data Centers must establish plans for the permanent archiving of the global long-term climate data record. The long-term archiving strategy must recognize the challenges imposed by the evolution of technology.

Long-term stewardship of data and products by the National Data Centers includes preserving and ensuring the accessibility of the data and products and their documentation. The science community has offered advice on the stewardship of Earth science data in long-term archives through the USGCR and CES. These are summarized from the NASA/NOAA Long-Term Archiving Plan:

- It is essential that the Long-Term Archive perform integrity checks on archive media between data migrations.

- Preservation and maintenance of data holdings including ensuring integrity and quality of the data and associated documentation is an essential function of the Long-Term Archive.

- It is essential that the Long-Term Archive develop and maintain a multi-year data migration plan.

- The Long-Term Archive should Migrate data sets to new, computer-compatible media on a regular basis, such that data sets are refreshed every 2 to 3 years with the pace of technology evolution.

- The Long-Term Archive should archive information on sensor development, calibration, and operating information, metadata and ancillary fields, operation product validation, and the basic radiances in a manner that allows reprocessing of Climate Data Records.

- The Long-Term Archive should evaluate on a regular basis the organization of data sets in the archive in light of actual user access and data usage patterns to improve efficiency of data access.

- The Long-Term Archive should develop flexible standards and formats that allow new services to be developed in the future.

- It is essential that the Long-Term Archive provide the next and subsequent generations of scientists with appropriate access to, and facilitate use of, its holdings.

- It is essential that the Long-Term Archive provide data and information services that are responsive to the needs of its users.

Data Access

The National Data Centers must provide timely access to their data holdings.

The Data Centers must adopt compatible metadata and interoperability standards that ensure that all data are accessible to the user community. The Data Centers must implement community-accepted tools for data access and utilization to provide the most efficient access to the data holdings.

The basic need of the global change researcher is accessibility to a dataset. Data customers must have easy access to the data either electronically or on media. An understanding of customers access to hardware, media devices and internet access is necessary for the data center to provide rapid data delivery at minimal cost.

Access to documentation and frequently asked questions (FAQ) will generally provide a data customer with the information needed to accurately access the dataset and understand the data's legacy. Data centers must provide easy to

find and current documentation to best serve their customers and to reduce the interaction time with staff.

Data centers must also provide and make easily accessible subsetting tools, format conversion tools, data manipulation and visualization tools to be able to serve a broader community.

The User Services Offices are the first interaction a customer experiences at a data center. The User Services Offices must be staffed with professionals experienced in Earth science disciplines. It is essential to the success of a data center to provide its customers access to a User Services Office that will be able to receive and track inquiries in a timely and thorough manner.

The User Services Office staff must be tasked to record information about users and their inquiries, following inquiries and orders for data to resolution, providing referrals, representing the end-user perspective in development efforts, acting as a conduit for user feedback and informing the user community of new products and services. The User Services Office is the advocate of the data customer and is the liaison between customers and data center personnel.

The User Services Office staff must also maintain the overall process of a Help Desk. Providing a mechanism via the web, email, telephone, fax and walk-in, to adequately establish clear communication with a customer.

Data Distribution

The National Data Centers must provide timely access to their data holdings.

The Data Center standards of practice must maximize performance while minimizing cost. To accomplish this, the Data Centers must take maximum advantage of the growing capacity of the Internet to support electronic delivery to users. The Data Centers can maximize data set usability by adopting and implementing standard practices for data subsetting, format conversion options, media options and availability of tools. Where cost effective, the Data Centers must develop and integrate new tools for the selective retrieval of science data and metadata attributes at the data server in preparation for distribution.

To provide an efficient data distribution system, the most cost-effective physical media must also be used as they gain acceptance in the user community.

To be effective, the Data Centers must keep current with distribution standards, including data compression, to reduce distribution volume. The Data Centers must ensure that standards of practice for distribution formats conform to Federal requirements (FGDC) as well as user preferences.

User Services

The National Data Centers must provide effective services to the user community.

Service is value added to the data. Forethought, careful planning, connection to the community's needs, anticipation of trends in technology and science will provide value to the data.

The Data Center User Services Offices are the front line of contact with the user community. Their effective performance, including support for users who require data and services from more than one Data Center, is essential for the success of the Data Centers. The primary responsibilities of the User Services Office include recording information about users and their inquiries, following inquiries and orders for data to resolution, providing referrals, representing the end-user perspective in development efforts, acting as a conduit for user feedback, and informing the user community of new products and services. The User Services Office documents new or alternate uses of data and products, and provides feedback, recommending data and product enhancements to facilitate alternate uses.

The Data Centers must develop data engineering and staff science expertise to provide technical and scientific information regarding the use of the data holdings. The Data Centers must encourage staff to invest time in attending community meetings, workshops and conferences to hear first hand of the activities, problems and needs of their user communities. The Data Centers must engage in active dialog with the user community to better understand their evolving data and service needs, and must improve services in response to community criticism and recommendations.

The Data Centers must define standardized levels of support to the user community and associated metrics to measure that support. The Data Centers must also define appropriate levels of service for each component of the user community, and collect and report statistics measuring that service.

The Data Centers must conduct periodic surveys of scientific user needs, of the needs of the flight projects, and of the needs of the supporting programmatic elements. The Data Centers will monitor measures of improved levels of customer satisfaction. Survey results are intended to provide context and perspective for the results of advisory group review. In addition, the National Data Centers must implement and monitor measures of user satisfaction with user interface / interoperability. They must also track the number and recognition of services provided and analyze user statistics for measures of user satisfaction.

The Data Centers must initiate a periodic peer review of their data management and stewardship, and of operations practices, standards, and performance.

Data Disposition

The National Data Centers must provide affordable permanent access to the data holdings.

The National Data Centers must prioritize their data holdings to permit the permanent archiving of those data that are required by the scientific community for the analysis of the global long-term climate record. The role of the scientific community in this prioritization must be recognized and made an integral part of the data disposition process.

Resource constraints may require that some data sets are relegated to passive but safe storage, and that some data sets are eliminated from the inventory. In both cases, the Data Center must ensure that science priorities guide any steps that are taken.

The Data Centers must develop a NARA-compliant data disposition plan for those data sets deemed unneeded or of very low priority by science advice. The data disposition plan may include transferring a data set to another appropriate agency that agrees to assume responsibility for it, or transferring responsibility to NARA.

References

1. “Data Management for Global Change Research Policy Statements”, The Office of Science and Technology Policy, D. Allan Bromley, July, 1991.

2. “Review of NASA’s Distributed Active Archive Centers”, Committee on Geophysical and Environmental Data, National Research Council, National Academy Press, 1998.

3. NASA Earth Science Strategic Enterprise Plan

4. ESE Statement on Data Management - 1998

5. 1999 EOS Reference Handbook

6. EOS Science Data Plan – November, 1998

7. EOSDIS DAAC Strategic / Management Plan – November, 1997

8. NASA NMI 8000.3 – March 22, 1991

9. NASA Performance Plan – February, 1999